Design and Application of Intelligent Visual Communication System for User Experience

Chao Peng
School of Art and Design, Henan Industry and Trade Vocational College, Zhengzhou, Henan 450000, China

***Abstract*—The design and application of visual communication system should be human-oriented, but currently this is often ignored by designers, resulting in poor user experience of visual communication. In order to improve the experience effect of visual communication system, combined with the existing computer technology, this paper proposes an intelligent visual communication system for user experience. First, for the problem of extracting multimodal features of users, considering the characteristics of different modal data, long and short-term memory networks are used to extract features with contextual information, and multi-scale convolutional neural networks are used for visual modality to extract low-level features from video frames. In the cross-modal stage, the low-level features in the source modality are used to enhance the target modality features. Then, for the personalized recommendation problem of users, a graph information extractor is constructed based on the graph convolutional neural network to fuse the recommended user-item bipartite graph node neighborhood information and generate a dense vector representation of nodes, which can enhance the recommendation effect in the form of incorporating the graph information representation in the deep recommendation model with Transformer as the sequence feature extractor. The proposed method is experimentally validated to shorten the response time and improve the performance of the system, which can increase the user experience of the visual communication system. The system designed in this article is user experience oriented, combined with Multimodal Features and intelligent recommendation algorithms, effectively meeting the personalized needs of users and has certain practical significance.**

Keywords*—Visual communication system; user experience-oriented; multimodal features; recommendation algorithm

I. Introduction

Visual communication was first used in the 1920s and refers to a design activity that conveys visual information by means of all visual media. Visual communication design is defined as a design that uses visual symbols to communicate information. Visual communication is included in design, which is also a purposeful and creative human practice. However, visual communication is different from other designs in that its primary function is to transmit information, which is different from product design and environmental design, where the use of perception is the main focus [1]. It is different from the transmission of abstract concepts that is done by language, and it involves a wide range of fields.

Visual communication enables the exchange of ideas and information between individuals and individuals. Vision is a physiological term. Vision is created by the action of light, the cells in the visual organ become active and excited, and the external visual content is processed by the visual nervous system to form vision. Through vision, human beings can perceive the size of external objects, the color of light and dark, color and the movement and stillness of objects, and in the process of perception, they can obtain various types of information that are important for the existence of the body. Vision is the most important of the five senses of human perception, and at least 80% of natural information in the natural environment is obtained through vision.

From the external form of visual communication, any act of communication must be carried out by means of physical symbolic media loaded with information, and these media must be visible. In the past, the most important medium of visual communication design was print, and the form of bearing was mainly two-dimensional plane [2].

Visual communication design is composed of three basic elements: text, logo and illustration. In order to communicate their thoughts and feelings and necessary information, humans have gradually developed language. The medium of spoken language is the beginning of the transformation and development of our communication behavior and the most fundamental medium in the progress of our social communication activities [3]. However, oral language has certain limitations. The first is the limitation of distance, because dialogue can only exist in close communication; the second is the constraint of time, because the oral language is not easy to record and store since there is no more export.

The visual communication system is a way for people to communicate through seemingly, using visual language to spread information. And the application of computer technology is to retouch and process the pictures. Nowadays, although people's living standards are improving, many people are busy with their lives and rarely have time to enjoy various scenery. In this case, the advantages of the visual tradition system emerge [4]. People do not have to go out to enjoy the scenery of various regions, which not only saves people's time but also satisfies their diverse visual experience. The relationship between visual communication among people, nature and society is shown in Fig. 1.

The visual communication system is an art form created by treating text, pictures, and colors as basic elements, and using advanced computer graphics software to process these elements so as to achieve a good artistic presentation effect, transforming a single performance into a more acceptable and favorite art form [5]. Nowadays, professional graphic image processing software such as Photoshop software, CAD series. This software will be based on the needs of visual communication design, text and other elements of the processing, to achieve a good visual communication effect.

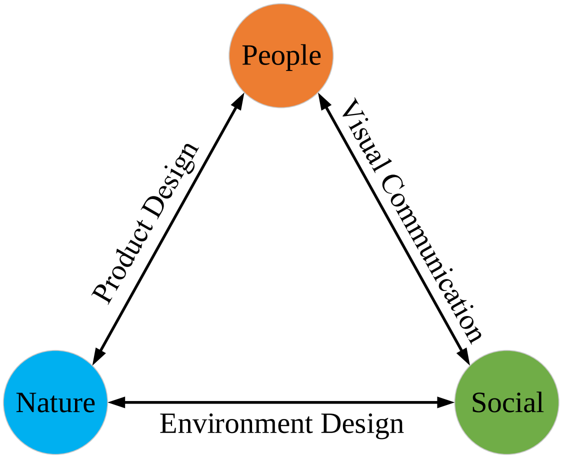

Fig. 1. Visual communication in the relationship between human, nature and society.

The application of these calculations in the visual communication system is discussed below: 1) Packaging design; 2) Advertising interface design; 3) Image processing technology. Nowadays, people like to use cell phones to take pictures, and in the process of taking pictures, the role of graphic image processing technology should not be underestimated, such as PS technology [6].

Computer technology is a way to use computer technology to modify, adjust and present. To achieve its function, it is necessary to build a specific scene of graphic description and create virtual optical effects with the help of light simulation technology, which is very closely related to computer geometry design. For computer graphics image processing technology, its data information is mainly from the subjective and objective world, the focus of image processing is to enhance the presentation of graphic images, which includes geometric image transformation, texture enhancement, color conversion and so on. At the same time, computer processing is based on digital signals, and it is widely used in the fields of medicine and aviation.

Based on the above background, this paper proposes an intelligent visual communication system for user experience in order to better improve the experience of visual communication system and combine with computer technology. Firstly, for the problem of multimodal feature extraction of users, an emotion recognition algorithm model with multimodal feature fusion is proposed, and each module and component of the proposed network model is introduced in detail. Then, for the user personalized recommendation problem, a new model architecture is designed by combining graph neural network with Transformer model to improve the recall of recommendation matching model. It is experimentally verified that the method in this paper has advantages in speed and performance.

The main research content of this article is as follows: First, for the problem of extracting multimodal features of users, considering the characteristics of different modal data, long and short-term memory networks are used to extract features with contextual information, and multi-scale convolutional neural networks are used for visual modality to extract low-level features from video frames. In the cross-modal stage, the low-level features in the source modality are used to enhance the target modality features. Then, for the personalized recommendation problem of users, a graph information extractor is constructed based on the graph convolutional neural network to fuse the recommended user-item bipartite graph node neighborhood information and generate a dense vector representation of nodes, which can enhance the recommendation effect in the form of incorporating the graph information representation in the deep recommendation model with Transformer as the sequence feature extractor.

II. Related Work

A. The Development Status of Visual Communication System

Visual communication system usually refers to the sender of information using visual symbolic elements to convey various information to the audience of information design, referred to as visual design visual communication includes two basic concepts, visual symbolic elements refers to what our eyes can see and can express the certain nature of things symbolic elements [7]. Visual communication design is also known as graphic design. The expressive design that we convey to the viewer's eyes and then shape the relevant content is collectively called visual communication design.

At this stage of analysis, the main content of visual communication design is still graphic design, professionals used to call it graphic design, visual communication design [8]. There is no great difference in the scope of design between graphic design and graphic design at this stage, and the relationship between them is a progressive one in terms of conceptual scope, not a contradictory and opposing relationship.

The study of non-planar elements of visual communication design is to break through the purely one-dimensional space to think and express around graphics, text and layout, and to advocate that all factors related to our design works should be included in the designer's scope of thinking and expression [9]. Visual communication design in the new era is not only about human vision, but also involves the sense of hearing and even touch.

In the context of the rapid development of computer media, visual communication technology, with its unique and innovative design images and language, has been gradually applied to the field of computer media, contributing to the development of the media. Visual communication technology is a kind of design activity that transmits and presents information to people through visual symbols, which are mostly expressed in the form of images [10]. Therefore, image design and processing is of great value in visual communication analysis, especially image fusion technology, which is very important for clear and complete presentation of visual images.

Computer graphic image processing techniques and visual communication systems are difficult to divide using standard lines in many aspects. The commonality between the two is reflected in the basic professional curriculum, color and pattern design, which can be said to converge between them. The aesthetics, design methods, design concepts and development backgrounds are the same, and there are also similar concepts in the arrangement of lines and surfaces [11]. In addition, the design techniques produced by both are based on the same design techniques, use and depth of exploration. In this condition, the interpolation of the two can be better realized.

Many visual image fusion related research results have appeared. Some researchers use visual weight map for image fusion, decompose multi-scale images using cross bilateral filters, calculate visual weight values at different decomposition layers, and complete image fusion by integrating the results of weight value calculation, but the quality of fusion still needs further improvement. Recently, some researchers use adaptive PCNN to extract image information and fuse images by inverse NSCT, and use different fusion strategies to fuse images several times to increase the fusion accuracy, which has better fusion quality but longer computation process [12]. There are also researchers who enhance the infrared image based on the guiding filter, inject the infrared image information into the visible image effectively, complete the image fusion, and post-process the image to enhance the fusion effect, the process is more complex and time-consuming.

At this stage, the society is developing in the direction of diversification, and in the future, the use of computer processing technology and visual communication design will gradually expand, it is necessary to discuss and study the problems of technology application in depth, and under this condition, propose effective solutions to promote the application of technology to obtain a better visual communication effect.

B. Application of Multimodal Features in Recommendation Algorithm

The essence of recommendation algorithm is to connect users and items in a certain way, and search and recommendation should be divided into at least two stages: recall and sorting. In the recall phase, because of the large amount of data processed, fast speed, simple models and few features are required.

Due to the need of business scenarios, online evaluation is likely to be the evaluation of multi-objective fusion results. The evaluation of offline experiments on search ranking focuses on two aspects: efficiency evaluation and effectiveness evaluation. Efficiency evaluates response time and space consumption. Effectiveness is comparing the ranking results with the standard results. The evaluation metrics include: accuracy, recall, F-value, average accuracy [13]. The accuracy of the system can be evaluated according to three different paradigms. The first is that the recommendation algorithm can be used as a rating prediction model, predicting the user's rating for all the subject matter that did not produce an action. The second is to view the recommendation algorithm as a classification problem.

All of these algorithms currently have the following drawbacks: the recommended items are very similar to the items previously purchased by the user, the recommended results can only cover a small portion of the items, and they cannot explore the user's changing interests [14]. With the increasing popularity of posting contents on social platforms, people will post images, texts and videos on microblogs and other platforms to exchange their thoughts and express their emotions.

Based on this scenario, we propose to extract information from multiple modalities to give personalized recommendations through sentiment computing. A novel DNN approach is proposed to exploit the features and class relationships in video classification by fusing speech, video, and text modalities using a joint architecture for video classification. The classification performance is improved by imposing regularization based on the trajectory paradigm on the specially tailored fusion and output layers and exploiting the commonality shared among semantic classes, however, the feature representation and classification models are not learned jointly [15].

Multimodal learning refers to learning information from each of the multiple modalities and enabling the exchange and transformation of information from each modality. Multimodal deep learning refers to building neural network models that can perform multimodal learning tasks. The prevalence of multimodal learning and the heat of deep learning have given multimodal deep learning a vivid vitality and development potential. The fusion architectures are divided into joint architectures, coordination architectures and codec architectures according to the different ways of feature fusion. The fusion methods include three model-independent methods, early, late and hybrid, and two model-independent methods, multi-core learning and image model. Modal alignment has been the difficulty of multimodal fusion technology, and the two commonly used methods are display alignment and implicit alignment.

As a technique that enables machines to possess more human intelligence, multimodal deep learning is expected to gain significant momentum in the future. The next step can be to further investigate the insufficiently researched issues such as semantic conflict of modalities, multimodal combination evaluation criteria, and modal generalization ability, to deeply explore the difficult issues such as cross-modal transfer learning and non-convex optimization, and to promote the application of this technology in some new areas of deep learning [16]. In order to better capture the inter-modal relationships, this paper proposes to adopt a model-independent approach. A weighted sentiment feature fusion model based on is proposed, thus providing support for the later calculation of sentiment similarity.

Based on the above analysis, it can be concluded that traditional methods suffer from poor clarity after image reconstruction or excessive redundancy in visual information feature extraction, resulting in unsatisfactory extraction results, low extraction efficiency, and long training time. Therefore, considering the characteristics of different modal data, this paper uses long short-term memory networks to extract features with contextual information, and uses multi-scale convolutional neural networks to extract low-level features from video frames. In the cross modal phase, low-level features in the source modality are used to enhance the target modality features, Integrating graph information representation into deep recommendation models to enhance recommendation performance. The experimental verification shows that the proposed method can shorten response time, improve system performance, and enhance the user experience of visual communication systems.

III. Algorithm Design

A. User Multimodal Feature Extraction

The model proposed in this section implements feature fusion of three modalities: image, acoustic and text. For the features of different modalities, a suitable deep neural network is selected for low-level feature extraction, and then the cross-modal attention mechanism proposed in this paper is used for feature fusion.

With the progress of machine learning and deep learning, machines that can achieve learning cognition and express emotions in complex real-world environments are often more versatile and effective in human-computer interaction-related research in artificial intelligence applications, which makes the role of affective computing more and more important. The purpose of this paper is to learn effective modal representations for feature fusion of textual, acoustic and visual multimodal data by using a cross-modal attention mechanism approach. The model structure of this paper is shown in Fig. 2.

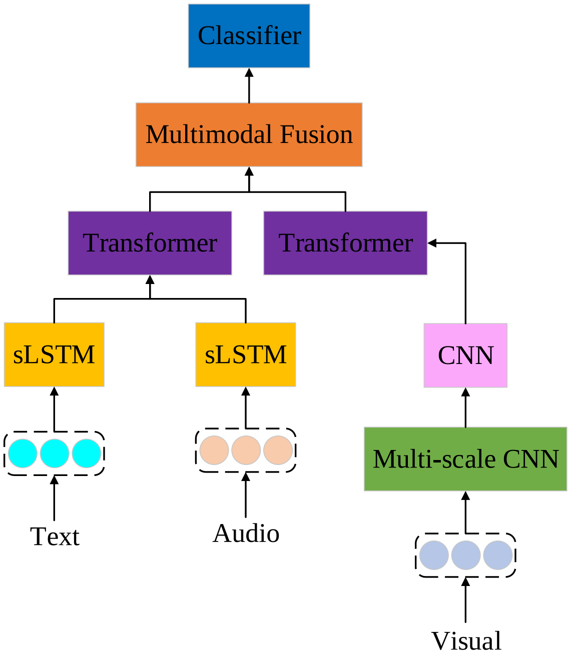

Fig. 2. Multi-modal feature fusion model.

The algorithm process is divided into four steps:

Step 1. In this paper, we use multimodal signals to detect emotions in videos, and use three modal data as the input of the model. The inputs include three low-level feature sequences from transcribed text, acoustic and visual modalities.

Step 2. Given the different characteristics of each modal feature, different neural networks are used to extract the features.

Step 3. After getting the different modal feature vectors, they are divided into two groups, text-acoustic and acoustic-visual.

Step 4. The output results of the two Transformer networks are spliced to obtain the fused features, and finally the prediction results are obtained through the fully connected network.

The advantages of LSTM networks in processing temporal data are: global processing and memory units. By processing data with time- or space-dependent characteristics, global information containing context-dependent information is obtained, and memory units are used to control the retention and removal of key historical information, enabling long- and short-term memory functions [17]. Therefore, it is more reasonable to use LSTM networks to capture the feature information of contextual cues. The textual feature vector obtained by using a one-way long and short-term memory network to capture temporal features, is calculated as follows:

$$t_i = sLSTM\left(T_i; \theta_i^{lstm}\right) \tag{1}$$

Since speech is highly random and correlated between adjacent frames, it is mainly reflected in the phenomenon of co-articulation when speaking, where words with connections before and after have an impact on the currently spoken word. Therefore, in this paper, for acoustic modality, the same LSTM network is used for feature extraction, and the acoustic feature vector is obtained and calculated as follows:

$$a_i = sLSTM\left(A_i; \theta_i^{lstm}\right) \tag{2}$$

For visual features, it is necessary to segment the video to obtain each video frame containing key feature information, sample it, and then use the face images in all video frames as the output information of acoustic modality. Therefore, for the visual modality considering the size of the input image data and the performance of the experimental equipment in this paper, MSCNN is applied and a set of feature maps are obtained, calculated as:

$$G_a^{MSCNN}\left(V, \theta\right) = \left\{v_a \middle| a \in \Omega\right\} \tag{3}$$

Where V denotes the input features.

For the different characteristics of the dataset, the first two are regression tasks, i.e., dichotomous tasks, and the commonly used loss functions are L1 Loss (Mean absolute loss, MAE) and L2 Loss (Mean Square Error, MSE). The MAE is calculated with absolute error as the distance, and is computed as shown below:

$$MAE = \frac{1}{n}\sum_{1}^{n}\left|y_i - \hat{y}_i\right| \tag{4}$$

MSE is also often used as a regular term, but because of the presence of the squared term, the gradient tends to explode when the predicted value differs significantly from the target value. Therefore, this paper uses MAE as the loss function and also introduces a varying learning rate.

B. User Recommendation Algorithm Design

Recommendation system is a research direction closely integrated with industrial applications, and the research of recommendation system from the academic point of view has certain limitations, therefore, the research on it needs to focus on the sub-problems of the recommendation task, or simplify the complex scenario, and optimize the sub-structure of the whole system from a certain point of view.

With the development of new neural network structures and efficient information extractors in recent years, the problem is solved by introducing graphical neural networks and Attention mechanism. Therefore, this chapter will introduce recommendation models based on graphical neural networks and Transformer structures, and show the effect of their application on classical datasets of recommendation systems and comparative experiments [18]. In addition, although the timestamps of interactions are introduced in the training process to constitute sequential information, it is necessary to distinguish this approach from intra-session recommendation, where individual sessions occur for a short period of time, while this model faces a long time span of data, which is still essentially using relatively static data for matching user preferences and item traits.

The core of the recommendation model proposed in this section is shown in Fig. 3, which incorporates information from the graph perspective of user-item interaction and applies an efficient feature extractor Transformer structure in order to obtain an effective method for item characterization with certain industrial practical capabilities.

Compared with graph embedding methods, the biggest advantage of graph neural networks in recommender system construction is their ability to be trained as part of an end-to-end model for collaborative filtering based on the introduction of graph structure. It is possible to use both node information and graph structure information.

The way node information is updated in a graph neural network is divided into two parts, i.e., message construction and message delivery. In each layer of GNN, messages can be passed along the edges of the graph. Each node receives messages from its neighbors and uses appropriate aggregation and mapping functions to construct new messages, which are used as node information in the next layer and also represent the new node features aggregated to the neighboring nodes' information. The message construction part mainly consists of the F mapping function, and the Sigma aggregation function. The Sigma function can be replaced by any aggregation function, and the F mapping function can be set as a single layer neural network:

$$M_i^r = F\left(\sum M_u^{r-1}, M_i^{r-1}\right) \tag{5}$$

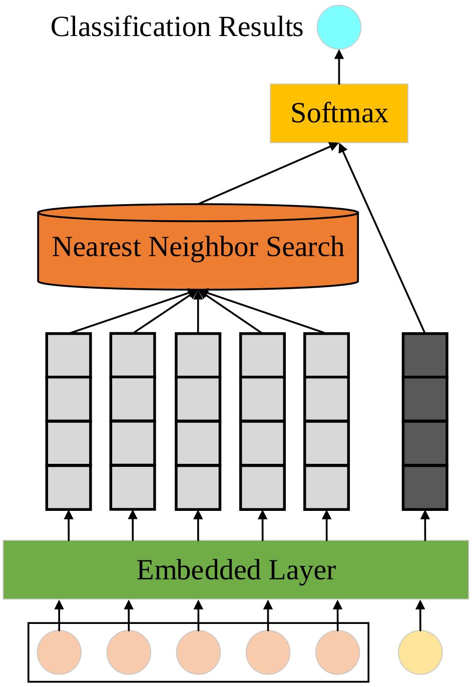

Fig. 3. Recommended model.

The difference between item-based and user-based recall algorithms lies in the pre-calculation of which of the two representations is used for nearest neighbor search. For most of the recommendation scenarios including movies and e-commerce, the recommended items are often relatively fixed, while the users are in constant high-frequency dynamic changes, and for large-scale e-commerce sites, the size of the users may be more than 10 times the size of the items.

The graph neural network structure used in the model is GCN, which uses the adjacency matrix, the degree matrix and the hidden feature matrix of the current layer to obtain the connection relationship between the nodes and the node features, and iteratively updates the node features:

$$H^{(l+1)} = Leaky\,\mathrm{Re}\,lu\left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(l)} W^{(l)}\right) \tag{6}$$

where the node hidden feature matrix H is the X-feature matrix in layer 0, the input layer.

In the training phase, the integrated features will be interacted with the candidate items, also embedded via items, in order to train the model parameters through a binary classification task. In the online service phase of the matching model, this feature tensor will be used as the retrieval target to obtain the matching result set by performing a K-NN nearest neighbor search from the full set of candidate items.

IV. Experiments

A. Experiment of Feature Extraction

Simulating pulse neural networks using computers requires powerful computing power. This article relies on the hardware devices listed in Table I for experiments:

TABLE I. HARDWARE TABLE OF DEVELOPMENT ENVIRONMENT

Development Environment	Parameter
GPU	V100S 32GB x 4
CPU	Intel(R) Xeon(R) Gold 6240R
Memory	38GB
Disk	7TB

The experiments first require setting hyper-parameters based on previous experience, and then tuning the parameters. The trainable parameters are obtained by back propagation, and the hyper-parameters must be set before building the entire network architecture. Each hyper-parameter related to the network structure needs to be specified first. For hyper-parameters such as discard rate and number of heads in the cross-modal attention module, a basic grid search method is performed in this paper. When the verification performance reaches a steady state, the learning rate is decayed by a factor of 10, making it more likely that the model will converge to a locally optimal solution, and the experimental results are cross-validated by a factor of 10. Two metrics, accuracy and F1 score values, are used to evaluate the performance of the algorithm model.

In order to fully verify the performance of the model proposed in this paper, the baseline models for experimental comparison are representative and recent algorithmic models with excellent performance. The baseline models are basically as follows: 1) EF-LSTM: a traditional network model using early fusion network and late fusion. 2) RMFN: a recurrent multi-stage fusion network that decomposes the fusion problem into multiple stages, each stage focusing on a subset of multimodal signal on a subset of the signal to achieve specialized and efficient fusion. 3) MFM: A multimodal decomposition model optimized for the joint generation of discriminative objectives for multimodal data and labels.

Table II gives the results obtained from the baseline model and the model proposed in this paper in the dataset, where the higher values of evaluation metrics Acc and F1 scores are better, and the comparison with the evaluation metrics Acc and F1 scores of different baseline models is used to demonstrate the optimal performance of the multimodal feature fusion sentiment recognition model proposed in this paper. In order to display the comparison results more intuitively, the best data shown in each column are bolded. It can be obtained that the sample recognition accuracy reaches 91.38%, the accuracy reaches 88.64%, and the average accuracy reaches 87.18%. The recognition accuracy of happy, angry and neutral emotions is the highest.

TABLE II. COMPARISON OF FEATURE EXTRACTION EFFECT OF EACH MODEL

Model	Acc	F1	Average
EF-LSTM	85.26%	83.72%	82.57%
RMFN	87.49%	86.43%	83.49%
MFM	90.54%	86.81%	85.36%
Ours	91.38%	88.64%	87.18%

Taking the Last.FM and Movielens datasets as examples, the distribution of the number of model training interactions is shown in Fig. 4. Although the overall interaction data scale is above 10M, the interaction volume has an obvious long-tail effect, which leads to the fact that the interaction volume of most items cannot meet the training scale of the model, which is reflected in the training results, only a very small number of items can be fully trained in a short time, and the overall training speed is slow. Therefore, negative sampling was introduced to expand the data in the training. When the negative sampling ratio is set to 5, that is, whenever a positive interaction is put into the training set, 5 samples of items that have not been interacted with the user will be collected at the same time, and during training, these 5 negative samples will be processed negatively, which greatly increases the training efficiency.

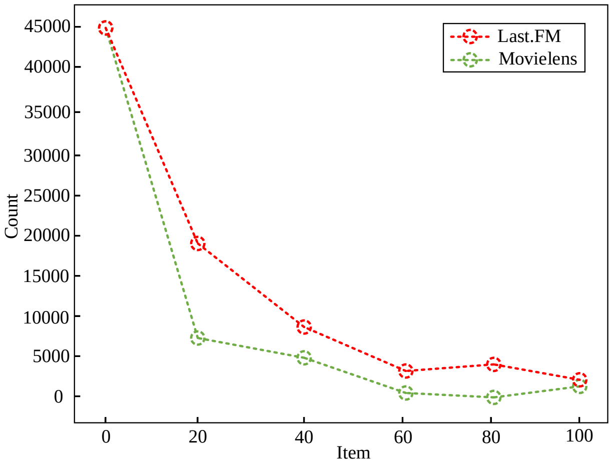

Fig. 4. Experimental results of model training interaction.

B. *Personalized Recommendation Experiments*

In the personalized recommendation experiments for users, the datasets used are: 1) Amazon-Music: containing 125,998 shopping reviews generated by 5,453 users and 65,833 items, with an interaction density of 0.35%. 2) Amazon-App: containing 341,784 shopping reviews generated by 18,328 users and 32,684 items, with an interaction density of 0.57%. The interaction density was 0.57%. 3) Movielens movie rating dataset, which contains user ratings of movies from the Movielens movie rating website collected by GroupLens.

The comparison methods are: 1) Popular+ClaSS(PopClass): a dynamic recommendation strategy based on popularity, which introduces the consideration of user preference categories. 2) DeepWalk: - a classical graph embedding technique, which constructs a sequence of nodes by random wandering on an undirected graph and applies (3) YoutubeDNN: In the matching phase, YoutubeDNN tries to use the information of items that users have interacted with and additional auxiliary information such as time information to input into the deep neural network as the user's representation, and then train the network through the multi-category classification task of videos. The steps can be broken down into two phases: network pre-training and model service.

The hit rate performance of our model and the comparison algorithm are shown in Table III. Since the Amazon-Music dataset is the sparsest, it indicates that there is incomplete training and therefore the overall hit rate is lower than the other datasets. In DeepWalk, the overfitting may be caused by the sparsity of the training data, but the graph-based embedding method is able to achieve the best hit rate performance based on a simple structure. However, the graph-based embedding method is able to achieve excellent performance based on a simple structure, which fully illustrates the importance of using the graph structure information in the field of recommendation recall.

In contrast to Transformer-Encoder, which also extracts sequence information, it should be noted that although this model achieves a good hit rate level, in industrial applications, the recall of target items by a single method only reflects a limited perspective of the recall strategy, and multiple fusion recall is needed in conjunction with business scenarios.

TABLE III. EXPERIMENTAL RESULTS OF HIT RATE OF EACH MODEL

Model	Amazon-Music	Amazon-App	Movielens
PopClass	13.25%	8.56%	4.52%
DeepWalk	24.63%	6.49%	12.61%
YoutubeDNN	22.84%	7.54%	10.84%
Ours	27.38%	11.27%	15.06%

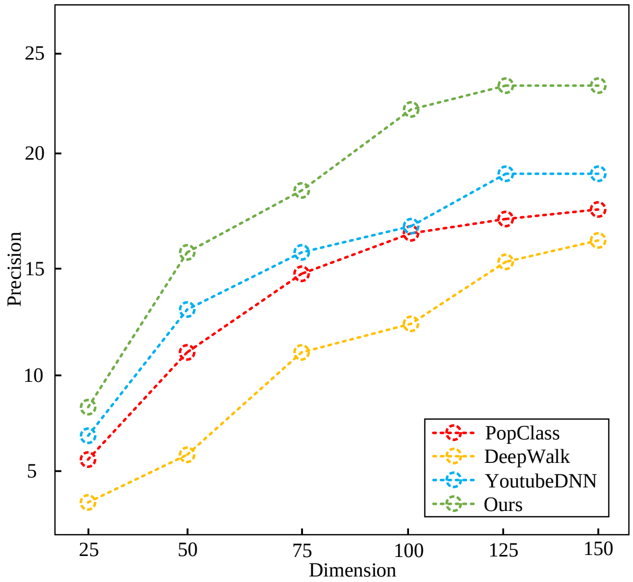

Fig. 5. Experimental results of each model under different implicit vector dimensions.

The experimental results of each model with different implied vector dimensions are shown in Fig. 5. It can be seen from the figure that the larger the implied vector dimension of the model, the better the performance of the model, because the larger the implied vector dimension, the more information the nodes contain, the finer the granularity of the representation of the user's interest, and thus the deeper and finer the granularity of the user's interest preference can be captured, and the recommendation effect will be better. However, due to the nature of the neural network structure, the implicit vector dimension increases the time and space complexity of the model, so the implicit vector dimension should be set to balance performance and resource consumption.

V. CONCLUSION

In order to improve user experience, this paper proposes an intelligent visual communication system oriented to user experience, and applies user experience to the algorithm design of the visual communication system to realize the design and application of personalized visual communication. First, for the user multimodal feature extraction problem, based on the cross-modal attention mechanism in its module, the target modal features are enhanced by including low-level features from other modalities. The output of multiple attention heads in it enhances the representational capability of the network and further better improves the effect of modal fusion. Then, for the personalized recommendation problem, the Transformer is used as the main structure of the sequence feature for extractor to replace the DNN in the base model, and the graph convolutional neural network module with end-to-end graph learning capability is combined with the dense vector representation of items in the recommendation matching stage to improve the hit rate of the matching algorithm. It is experimentally verified that the method in this paper achieves a large performance improvement on three different datasets compared with other advanced network models.

Although convolutional neural networks have developed relatively maturely in the field of visual recognition, capturing subtle changes in visual images that are difficult to detect, efficiently extracting facial expression features, and effectively utilizing these features are key issues; Meanwhile, how to fully consider the changes in key facial features, as well as the fusion of global and local features, is also a major key factor in the process of facial expression recognition. Although the paper proposes a visual feature recognition method based on convolutional neural networks to address the above issues, which has improved the recognition accuracy to some extent, our work still has some shortcomings and needs to be improved and perfected in the following aspects in the future.

Firstly, for static image recognition, due to the limitations of the dataset itself, the accuracy of recognizing certain expression categories is still lacking. Therefore, future work will focus on training and validating the proposed method on datasets with more evenly distributed expression labels.

Secondly, the method proposed in this article has not yet been applied to datasets collected in real environments. Future work will train and validate the method using datasets collected in real environments.

Finally, for the local collection end, when performing facial expression recognition, only one person's facial image can be recognized, and the problem of storing expressions when multiple people appear simultaneously has not been solved; For online management, we hope to achieve real-time facial expression collection on web pages in the future to reduce costs.

REFERENCES

[1] Alhayani, B. S., & Llhan, H. (2021). RETRACTED ARTICLE: Visual sensor intelligent module based image transmission in industrial manufacturing for monitoring and manipulation problems. Journal of Intelligent Manufacturing, 32(2), 597-610.

[2] Tan, J. K., & Sato, A. (2020). Human-robot cooperation based on visual communication. International Journal of Innovative Computing, Information and Control, 16(2), 543-554.

[3] Hassan, M. K., Hassan, M. R., Ahmed, M. T., Sabbir, M. S. A., Ahmed, M. S., & Biswas, M. (2021). A survey on an intelligent system for persons with visual disabilities. Aust. J. Eng. Innov. Technol, 3(6), 97-118.

[4] Ma, C., Li, X., Li, Y., Tian, X., Wang, Y., Kim, H., & Serikawa, S. (2021). Visual information processing for deep-sea visual monitoring system. Cognitive Robotics, 1(2), 3-11.

[5] Kountouris, M., & Pappas, N. (2021). Semantics-empowered communication for networked intelligent systems. IEEE Communications Magazine, 59(6), 96-102.

[6] Wang, B., Xu, K., Zheng, S., Zhou, H., & Liu, Y. (2022). A deep learning-based intelligent receiver for improving the reliability of the MIMO wireless communication system. IEEE Transactions on Reliability, 71(2), 1104-1115.

[7] Yang, J., Wang, C., Jiang, B., Song, H., & Meng, Q. (2020). Visual perception enabled industry intelligence: state of the art, challenges and prospects. IEEE Transactions on Industrial Informatics, 17(3), 2204-2219.

[8] Liu, R. W., Guo, Y., Lu, Y., Chui, K. T., & Gupta, B. B. (2022). Deep network-enabled haze visibility enhancement for visual IoT-driven intelligent transportation systems. IEEE Transactions on Industrial Informatics, 19(2), 1581-1591.

[9] Lan, Q., Wen, D., Zhang, Z., Zeng, Q., Chen, X., Popovski, P., & Huang, K. (2021). What is semantic communication? A view on conveying meaning in the era of machine intelligence. Journal of Communications and Information Networks, 6(4), 336-371.

[10] Dodda, S., Chintala, S., Kanungo, S., Adedoja, T., & Sharma, S. (2024). Exploring AI-driven Innovations in Image Communication Systems for Enhanced Medical Imaging Applications. Journal of Electrical Systems, 20(3s), 949-959.

[11] Imoize, A. L., Adedeji, O., Tandiya, N., & Shetty, S. (2021). 6G enabled smart infrastructure for sustainable society: Opportunities, challenges, and research roadmap. Sensors, 21(5), 1709-1719.

[12] Fu, Y., Li, C., Yu, F. R., Luan, T. H., & Zhang, Y. (2021). A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance. IEEE transactions on intelligent transportation systems, 23(7), 6142-6163.

[13] Alhayani, B., Abbas, S. T., Mohammed, H. J., & Mahajan, H. B. (2021). Intelligent secured two-way image transmission using corvus corone module over WSN. Wireless Personal Communications, 120(1), 665-700.

[14] Alhayani, B., Abbas, S. T., Mohammed, H. J., & Mahajan, H. B. (2021). Intelligent secured two-way image transmission using corvus corone module over WSN. Wireless Personal Communications, 120(1), 665-700.

[15] Liu, R. W., Nie, J., Garg, S., **ong, Z., Zhang, Y., & Hossain, M. S. (2020). Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6G-enabled maritime IoT systems. IEEE Internet of Things Journal, 8(7), 5374-5385.

[16] Njoku, J. N., Nwakanma, C. I., Amaizu, G. C., & Kim, D. S. (2023). Prospects and challenges of Metaverse application in data-driven intelligent transportation systems. IET Intelligent Transport Systems, 17(1), 1-21.

[17] Gao, X., Wang, Y., Chen, X., & Gao, S. (2021). Interface, interaction, and intelligence in generalized brain–computer interfaces. Trends in cognitive sciences, 25(8), 671-684.

[18] Lazaroiu, G., Androniceanu, A., Grecu, I., Grecu, G., & Neguriță, O. (2022). Artificial intelligence-based decision-making algorithms, Internet of Things sensing networks, and sustainable cyber-physical management systems in big data-driven cognitive manufacturing. Oeconomia Copernicana, 13(4), 1047-1080.